Short Read Archive (SRA) Submission Guidelines

National Center for Biotechnology Information (NCBI)
National Library of Medicine

25 June 2008 Version 0.8 Draft B

Contents

1 Overview

1.1 Scope

The Short Read Archive (SRA) at NCBI accepts primary sequencing data from "next generation" sequencing platforms, including Roche 454®, Illumina Illumina®, Applied Biosystems SOLiD®, Helicos Biosciences HeliScope®, and others.

Sequencing data should be submitted to the SRA rather than the regular Trace Archive. The Trace Archive is intended as the repository of sequencing data from gel/capillary platforms (Applied Biosystems 370® and 3730® , Megabace, and Licor sequencers).

The SRA can also accept sequence assemblies and reference alignments.

1.1.1 Provisional SRA

Currently NCBI operates a "provisional" version of the Short Read Archive (SRA) that accepts short read sequencing deposits, assigns permanent top level accessions to the submissions, incorporates metadata into xml format, and makes the run data available to the public in their original format. The goal of this resource is to provide basic services to the community for the archival of short read datasets.

1.1.2 Production SRA

All data submitted to the provisional SRA will migrate to the "production" SRA when the production version of this resource is deployed. The Production SRA will serve datasets through NCBI web servers with some search and browsing capability and high throughput submission support. Release of the Production SRA is scheduled for Feb 2008.

1.2 Related Documents

- NCBI Short Read Archive
- SRA Tracking Page
- SRA XML Schema Documents
- Sequence Read Format

1.3 Revision History

Author	Release	Notes
Martin Shumway	25 June 2008 Version 0.8 Draft B	Data on SOLiD, released for review by Zamin Iqbal
Martin Shumway	16 June 2008 Version 0.8 Draft A	Moved information about data transfer to companion document "High Throughput Primary Data Transfer to NCBI"
Martin Shumway	31 March 2008 Version 0.7 Draft A	Added formulas for producing SRF files from Illumina run folders
Martin Shumway	21 March 2008 Version 0.6 Draft H	Added ftp guidelines, more guidelines on obtaining accounts, exactly specify submission protocol
Martin Shumway	28 Feb 2008 Version 0.6 Draft G	Reviewed by Dennis Benson
Martin Shumway	25 Feb 2008 Version 0.5 Draft E	Draft distributed on SRA web site v0.6
Martin Shumway	24 Jan 2008 Version 0.5 Draft D	First draft shown to 1000Genomes project

2 Terms of Usage

2.1 Permanence

Accessions issued by the SRA are always maintained and never reused. If a desired record has been withdrawn, then a message to this effect will be displayed to anyone who tries to access it. If a record has been superseded by a successor record, this fact will be presented to anyone trying to access it. Only in rare cases where the record needs to be expunged from the archive will a user not be able to access it.

2.2 Authentication

Submissions are managed through secure channels. These channels include MyNCBI, NIH level login through CIT, and ftp accounts secured by passwords. We will correspond with submitters via email about submission and curation issues, but we do not exchange data by email.

Individuals may obtain a MyNCBI account for their transactions. Please do not reuse someone else's MyNCBI account. Center accounts are provided for the convenience of automated pipelines. The authentication information for such an account must be maintained securely by the Center. Accounts may be disabled or withdrawn after a long period of disuse in order to comply with NCBI security requirements.

2.3 Limitations

The Short Read Archive at NCBI is a public resource and the decision whether to submit data to this resource is the responsibility of the submitter. Prospective submitters should be aware of the following issues:

- Never submit data without the permission of the **principal investigator**.
- Most **human data** gathered from research subjects are under strict privacy controls and must be handled with privacy protections as determined by the research institution's Institutional Review Board (IRB), the funding agencies, and the laws of the United States or the submitter's home country. The dbGaP resource at NCBI may be a more appropriate broker for human sequencing data.
- Data submitted as part of a journal manuscript may have a **publication embargo** placed on it by the journal editors. The submitter can place a "hold until publish" restriction on the submission to the SRA as part of the submissions process.
- Data that might relate to **patents** and **intellectual property** may be submitted to NCBI, but the submitter is responsible for ensuring that procedures and policies of his/her institution or company are observed.
- Some **environmental data** gathered in the territory of certain countries, including territorial waters, may have sovereign legal restrictions on their use. NCBI cannot accept such data since NCBI is not able to enforce any usage restrictions..
- Submitters must ensure that data obtained as part of a criminal investigation is free of any judicial restrictions on its use.
- Submitters are responsible for obtaining any necessary permissions from the collecting institution for **forensic and paleontological data.**

2.4 Modification

NCBI allows submitters to modify their records. Such requests must be formally entered using the SRA submission mechanisms. Informal requests by email will not be accepted. Only the center or individual that created the record can change it. Please write NCBI if you have changed affiliations and wish to update old records. This may require agreement from the original institution.

2.5 Curation

From time to time records deposited at the SRA must be updated with changes needed in order that the data continue to conform with the data model for the archive, to update data as it changes (for example finalizing publication information), to change data that are clearly wrong (for example correcting external references to other data or resources), and to add additional relevant metadata as they become available. NCBI will contact the submission owner on a best effort basis. The submission owner should maintain up to date contact information with NCBI to receive word of such changes.

Actual instrument data are never changed by NCBI. Only the submitter can make such modifications.

2.6 Availability

While NCBI tries to maintain maximum uptime of its servers on a 24x7 basis, no guarantee of availability is offered to users. Submissions that are interrupted by downtime may have to be restarted by the user.

Technical assistance is available on a limited basis during business hours USA Eastern Time. There is no guarantee for level of service regarding manual assistance.

3 Data Model

The SRA data model is discussed in detail elsewhere, but here is a brief overview.

The SRA tracks the following five objects:

Study – Identifies the sequencing study or project and contains multiple experiments.
Sample – Identifies the organism, isolate, or individual being sequenced.
Experiment – Specifies the sample, sequencing protocol, sequencing platform, and data processing that will result one or more runs.
Run – Identifies run data files, the experiment they are contained in, and any runtime parameters gathered from the sequencing instrument.
Analysis – Packages data associated with short read objects that are intended for downstream usage or that otherwise needs an archival home. Examples include assemblies, alignments, spreadsheets, QC reports, and read lists.

In addition, all details concerning submissions are contained in a separate document called **Submission**, which contains center specific submitting information, contacts, actions for the archive, and a file manifest.

Objects can be archived in the SRA at different points in time. Multiple submissions documents can be submitted. For example, study, sample, and experiment objects can be created at an early stage, with run data being submitted as the data are produced.

All SRA objects that are being created with XML files can be referenced by an alias. This is even true after they have received an accession. The namespace that the alias must be unique in is that of the submitting center.

4 Communicating with NCBI

4.1 Obtaining NCBI Accounts

4.1.1 Establish a NCBI Identity

Before interacting with NCBI, please obtain a personal identity. This will allow you to make submissions, track results, change records now or later, and hold or release records. There are three kinds of identity each of which is sufficient to do business with NCBI:

- MyNCBI – User-created and managed account for NCBI users.
- NCBI PDA – NCBI-created and managed account for primary data submitters. If you belong to a submitting Center and will play a role in monitoring and maintaining primary data submissions, please identify this fact through your account profile and also email NCBI with this information.
- NIH – Any NIH personnel has credentials managed through CIT and can use their NIH identity to login

4.1.2 Establish a Center Name

If your dataset is large, or you plan on submitting more than once in the future, or you have an existing relationship with NCBI, then you should use a dedicated secure trace ftp account.

First check to see whether you can use an existing account for your center. You will also need to know what your center name is. Please consult the current Center Name list to look up your center name. If your center has been submitting traces in the past, then an ftp account should exist. Write us and we can direct you to the submission contact from your center.

If you do not have an institutional ftp account, please write sra@ncbi.nlm.nih.gov requesting one. You will need to provide institutional as well as personal contact information. The ftp account is secured by a password known only to yourself and NCBI. The password may be shared with colleagues who need to participate in submissions.

4.2 Submitting Data

4.2.1 High Throughput Submissions

High-volume submissions should be uploaded to the ftp directory for your center. To do this,

```
ftp ftp-trace.ncbi.nlm.nih.gov
login: myaccount_trc
passwd: !jXYZZ3@ce

  > cd short_read
  > put myfiles.tgz
  > quit
```

You should double check that the file size that you posted agrees with the original file.

You cannot delete files once they are posted. Please write to sra@ncbi.nlm.nih.gov to request deletions.

4.2.2 Individual Submissions

Individual submitters may submit files through anonymous ftp to NCBI. This account is shared by all such submitters, but the files once written are not readable even by the submitter. In this case the center name is "Individual" and submissions are not tracked by institution. Please ensure that contact information written into the submission documents will allow NCBI to contact you to confirm receipt of the files and to communicate any problems.

To submit individually, please

1. Create a MyNCBI account
2. Write sra@ncbi.nlm.nih.gov to request the ftp address of the current anonymous ftp box.
3. Put the file to the anonymous ftp box, for example *mysubmission.tar.gz*

The Individual Submissions channel is intended for small submissions or uploads of test submissions when the submitter does not yet have a private account.

4.2.3 Interactive Submissions

Soon a web tool will be provided that allows you to create a submission online, upload files, and manage all your submissions. To use this tool, you will connect to NCBI through one of your NCBI identities, then proceed to the submissions management tool.

4.3 Packaging Data

4.4 Transmitting Data

Please see the companion document "High Throughput Primary Data Transfer to NCBI" for ftp and ftp alternatives.

4.5 Tracking Submissions

The Short Read Archive Submissions page tracks submissions by SRA number and current status. A link is provided to the XML metadata documents and the submitted run time data on the outgoing ftp site.

This page will be replaced once the SRA has entered full production.

5 Creating New Submissions

This use case covers new submissions which have not yet received an accession.

5.1 Preparing Run Data

The SRA is intended as a repository of data output by "primary analysis" phase of the sequencing platform: sequencing results in fasta form along with instrument data indicating probability of correctness for each basecall (qualities) and signal intensity measurements (intensities).

The Short Read Archive does NOT accept fasta only datasets due to the inability to evaluate the quality of such data.

5.1.1 454

The SRA accepts deposits of short read data from the 454 platform in the *.sff* format. These files should reflect the sequencing run setup. If the entire picotitre plate was used, then one *.sff* file per run should be submitted. If on the other hand the picotitre plate was divided into two or more regions, then a *.sff* file for each region should be submitted. If a *.sff* file contains more than one run, or more than one region in the run, please break up this file into constituent parts using the *sfffile* utility from the "Off Rig" software package provided by Roche.

Data Series	Number of Channels	Description
.sff	1	Flowgram (base call, phred quality score, flow value)

The read names found in the *.sff* file are meaningful and reflect the addressing scheme for the picotitre plate as well as a globally unique run id. Please do not rewrite this name as such addressing information will be lost. The sff file format is nearly optimal in terms of

footprint, so there is little to be gained by further compressing them. Therefore, please provide *.sff* files uncompressed.

The sequencing data may have been produced by the 454 contract sequencing center (454MSC). Please ask 454MSC to provide *.sff* files for your project.

5.1.2 Illumina Genome Analyzer

Illumina data can be submitted at several levels of detail, depending on the goals of the submitter.

The conduit for Illumina primary data is the short read format (SRF). Users should download the Staden io_lib package in order to get the *solexa2srf* utility. As an alternative to SRF, NCBI currently supports "native file" format submission[1]. These should be organized into a compressed tape archive file (*.tar.gz*), with all the files from each lane constituting one tar file.

5.1.2.1 Illumina Primary Data for Alignment and Abundance Measurement

Sequencing data with minimal instrumentation output is appropriate for applications where the main goal is abundance measurement rather than reconstruction of original sequence.

Data Series	Number of Channels	Description
_seq.txt	1	Base calls per read
_prb.txt	4	Per channel log odds quality scores

To produce a primary analysis SRF submission file for a lane's worth of data, change the working directory to the run folder and do:

```
solexa2srf -R -P -N <run>:%l:%t: -n %x:%y
    -o <center_name>_<run>_<lane>.srf   s_<lane>_*_seq.txt
```

where <center_name> is the short name of the sequencing center or other individual name, <run> is the flowcell name for the run (for example `080117_EAS56_0068`), and <lane> is the desired lane.

To produce a primary analysis SRF submission file for a lane's worth of paired-ends data, change the working directory to the run folder and do:

```
solexa2srf -R -P -N <run>:%l:%t: -n %x:%y -2 <cycle>
    -o <center_name>_<run>_<lane>.srf   s_<lane>_*_seq.txt
```

[1] Software engineers should plan to support the common input formats such as SRF when designing automated submission systems.

where <center_name> is the short name of the sequencing center or other individual name, <run> is the flowcell name for the run, <lane> is the desired lane, and <cycle> indicates the cycle number that starts the second read.

5.1.2.2 Illumina Primary Data with Calibrated Qualities

The following formulation uses single-dimension quality scores calibrated for the type of data sequenced and transformed to the Phred probability-of-error model.

Data Series	Number of Channels	Description
_seq.txt	1	Base calls per read
_qcal.txt	1	Single channel phred quality scores (one file per mate pair for paired end libraries).

5.1.2.3 Illumina Primary Data for Filtering and SNP Discovery

Mid level instrument data output from "four-channel primary analysis" is appropriate for most use cases, including *de novo* sequencing and resequencing, structural variation discovery, and SNP calling.

The conduit for Illumina primary data is the short read format (SRF). Users should download the Staden io_lib package in order to get the *solexa2srf* utility. If the goal is to provide data that can be filtered for quality and displayed, then the "processed" (calibrated) data series should be included in the output.

Data Series	Number of Channels	Description
_seq.txt	1	Base calls per read
_prb.txt	4	Per channel log odds quality scores
_sig2.txt	4	Phase-corrected signal intensity values

To produce a primary analysis SRF submission file for a lane's worth of data, change the working directory to the run folder and do:

```
solexa2srf -N <run>:%l:%t: -n %x:%y
    -o <center_name>_<run>_<lane>.srf   s_<lane>_*_seq.txt
```

where <center_name> is the short name of the sequencing center or other individual name, <run> is the flowcell name for the run (for example `080117_EAS56_0068`), and <lane> is the desired lane.

To produce a primary analysis SRF submission file for a lane's worth of paired-ends data, change the working directory to the run folder and do:

```
solexa2srf -N <run>:%l:%t: -n %x:%y -2 <cycle>
    -o <center_name>_<run>_<lane>.srf   s_<lane>_*_seq.txt
```

where <center_name> is the short name of the sequencing center or other individual name, <run> is the flowcell name for the run, <lane> is the desired lane, and <cycle> indicates the cycle number that starts the second read.

At the option of the submitter, calibration lanes or reads, and reads failing purity and chastity measurement thresholds, may be omitted (*solexa2srf –c*).

Each flowcell contains 8 lanes but not all lanes are used for production. Also, some lanes are devoted to other projects. Finally, the size of the SRF file produced by this process can be expected to be about 2 GB. For these reasons, it is desirable to produce one SRF file per lane. The SRF file format is nearly optimal in terms of footprint, so there is nothing to be gained by further compressing them. Therefore, please provide *.srf* files uncompressed.

5.1.2.4 Illumina Primary Data for Reprocessing

Low level instrument data output from "four-channel primary analysis" is appropriate for most use cases requiring great sensitivity to instrumentation data, including SNP calling and structural variation analysis, and as proof of concept for novel sequencing techniques.

If the goal is to provide data that can be reprocessed by another center, then the "raw" (uncalibrated) data series should be included in the output.

Data Series	**Number of Channels**	**Description**
_int.txt	4	uncalibrated signal intensity values
_nse.txt	4	uncalibrated noise profiles
_seq.txt	1	Base calls per read
_prb.txt	4	Per channel log odds quality scores
_sig2.txt	4	Phase-corrected signal intensity values

To produce a primary analysis SRF submission file for a lane's worth of data, change the working directory to the run folder and do:

```
solexa2srf -r -p -N <run>:%l:%t: -n %x:%y
    -o <center_name>_<run>_<lane>.srf   s_<lane>_*_seq.txt
```

where <center_name> is the short name of the sequencing center or other individual name, <run> is the flowcell name for the run (for example `080117_EAS56_0068`), and <lane> is the desired lane.

To produce a primary analysis SRF submission file for a lane's worth of paired-ends data, change the working directory to the run folder and do:

```
solexa2srf -r -p -N <run>:%l:%t: -n %x:%y –2 <cycle>
    -o <center_name>_<run>_<lane>.srf   s_<lane>_*_seq.txt
```

where <center_name> is the short name of the sequencing center or other individual name, <run> is the flowcell name for the run, <lane> is the desired lane, and <cycle> indicates the cycle number that starts the second read.

All data in the flowcell should be included and data should not be filtered for chastity or purity.

Each flowcell contains 8 lanes but not all lanes may be used for production. Also, some lanes are devoted to other experiments. Finally, the size of the SRF file produced by this process can be expected to reach 10-12 GB. For these reasons, it is desirable to produce one SRF file per lane. The SRF file format is nearly optimal in terms of footprint, so there is little to be gained by further compressing them. Therefore, please provide *.srf* files uncompressed.

5.1.3 Applied Biosystems SOLiD System

Primary analysis data from the SOLiD System is delivered in "color space", without translation into base space. Quality scores and signal intensities are based on the color calls.

While SRF translation software is under development, NCBI currently supports "native file" format submission[2]. These should be organized into a compressed tape archive file (*.tar.gz*), with all the files from one run constituting one tar file.

5.1.3.1 SOLiD Primary Data for Alignment and Abundance Measurement

Sequencing data with minimal instrumentation output is appropriate for applications where the main goal is abundance measurement rather than reconstruction of original sequence.

Data Series	Number of Channels	Description
.csfasta	1	Base calls per read in color space
_QV.qual	1	Color space quality scores

For paired end data two files of each file type will exist (F3 and R3).

5.1.3.2 SOLiD Primary Data for Filtering and SNP Discovery

[2] Software engineers should plan to support the common input formats such as SRF when designing automated submission systems.

Mid level instrument data output from "four-color primary analysis" is appropriate for most use cases, including *de novo* sequencing and resequencing, structural variation discovery, and SNP calling.

Data Series	Number of Channels	Description
.csfasta	1	Base calls per read in color space
_QV.qual	1	Color space quality scores
_intensity.ScaledCY3.fasta	1	Color processed intensity measurement
_intensity.ScaledCY5.fasta	1	Color processed intensity measurement
_intensity.ScaledFTC.fasta	1	Color processed intensity measurement
_intensity.ScaledTXR.fasta	1	Color processed intensity measurement

For paired end data two files of each file type will exist (F3 and R3).

Obtaining processed intensity measurements from the SOLiD System primary data requires appropriate configuration of the sequencing instrument. Contact the vendor for details.

5.1.4 Helicos HeliScope

As of this writing too little information is available on file formats from the HeliScope platform. This document will be updated as such information becomes available. If you have data from these platforms, please write us at sra@ncbi.nlm.nih.gov for special instructions.

5.1.5 Run Descriptor

A run descriptor XML file packages all the run data. Please consult the schema for required structure and information.

Submission Object	Description	XML Schema specification
Run	One of more XML descriptors linking run data to their experiments.	SRA.run.xsd

5.2 Preparing Metadata

A salient feature of the SRA is the distinction given to metadata. Rather than embedding these with every run record, short read metadata are organized into a collection of XML files that capture as much, or as little, information as the submitter cares to give. Many pieces of information can be provided in the form of links and tag-value pairs, eliminating the need to negotiate complicated data representation ontologies.

Submission Object	Description	XML Schema specification
Study	XML file specifying sequencing study	SRA.study.xsd

Sample	XML file specifying the target of sequencing	SRA.sample.xsd
Experiment	XML file specifying experimental organization and parameters	SRA.experiment.xsd

5.2.1 Genome Project Registration

Whole genome sequencing projects should be registered with the Entrez Genome Projects resource at NCBI before submitting to SRA. Please access the Entrez Genome Project Submission Form to submit.

5.2.2 Taxonomy Registration

Most single organism genome and transcriptome sequencing projects need a Taxon Id to help specify the sample being sequenced. Please consult the Entrez Taxonomy resource to see whether your organism is represented, and request an entry to be created if not. The taxon id is needed for submission preparation.

5.2.3 Reference Fields and Namespaces

All the XML files can take either names (aliases) to identify dependencies. These names need only be unique throughout the submission. Eventually, the SRA will replace these names with actual accessions while preserving the referential integrity of the records.

5.2.4 Required Fields

Each XML file has certain required fields. The XML schema document these entries and most are self-explanatory. Decisions that the submitter should make include:

- Study type
- Whether a genome project id exists for the study
- The center project name or id
- Whether a taxon id exists for the sample
- Whether an anonymous id exists for the sample
- Sequencing platform used (for example, 454 GS 20 or 454 GS FLX)
- Library source, strategy, selection, layout if applicable, protocol
- Spot or cluster layout (use of adapters, linkers, bar codes, etc)
- Read processing selection

In addition, the submitter should think about the relationship between experiments and samples, runs and experiments, and whether to split any of these objects to represent distinct information

Finally, submitters may consider providing ancillary information, including links, Entrez links, and attribute tag-value pairs. These can be created for any of the five record types.

5.3 Preparing Analysis Files

The analysis package contains submitted files pertaining to auxiliary analysis (for example assemblies of sequencing data), or data intended for downstream archival.

These data receive accessions, but there is little structure to them. The files are served back to the public in their original formats.

Submission Object	Description	XML Schema specification
Analysis	Auxiliary or pass-through data	SRA.analysis.xsd

5.4 Preparing Submission Files

Aspects of the submission process pertaining to the submission itself have been broken out into their own XML descriptor. Contact information, transaction requests, exceptions, and file manifests can be listed here. Contacts should be provided for questions or problems pertaining to the particular submission.

Submission Object	Description	XML Schema specification
Submission	XML file specifying submission session	SRA.submission.xsd

A checksum should be computed for each run file delivered as part of the submission and entered into the *submission.xml* record. Please use the unix *md5sum* or equivalent utility. It is not necessary to provide checksums for the metadata xml files.

5.5 New Submission Protocol

1. Check your XML files for correctness with respect to the published schema, for example:

```
xmllint --schema \
http://www.ncbi.nlm.nih.gov/Traces/sra/static/SRA.study.xsd \
myStudy.xml
```

2. Check your XML files for completeness and referential integrity.
3. Verify checksums on run and analysis files.
4. Open *ftp* to the trace ftp site for your Center.
5. Change directory to *./short_reads*
6. Deposit the files using the ftp *put* or *mput* command. The files may be rolled into a single tar archive file.
7. Confirm receipt of the submission in the SRA Tracking Page. Once processed you will be able to download the submitted data through this page. Unfortunately, while the provisional SRA is in operation processing is manual and so we cannot guarantee any particular response time. Please let us know if you are under a tight publication deadline and we will try to accommodate your needs.
8. Please write to us at mailto:sra@ncbi.nlm.nih.gov with any questions about status or access.

6 Managing Existing Submissions

6.1 Update Submissions

Updates can be performed on the Project, Sample, and Experiment objects by submitting replacement XML files for the affected objects. Because they depend on content, run and analysis objects must be replaced outright (withdraw followed by add). An update XML document must identify the target accession.

6.2 Hold Until Publish

An essential feature of the SRA is the ability to hold a submission until a manuscript reporting on the research is accepted or released by a journal. There are three variants:

- Hold for number of days – This is appropriate for certain data release policies.
- Hold until date – This is appropriate for scheduled release of a publication
- Hold – This is appropriate when a publishing journal has not yet been determined, or the publication date has not yet been set.

The hold can expire its term, or the submitter may send a Release message to NCBI indicating that the submission can be released to the public. Minimal tracking information about the submission including accession, date, center, title, platform, and size statistics are displayed on the SRA tracking page regardless of hold status.

A Release message can apply to the entire SRA object, or individual objects within the SRA submission. Any dependent objects are implicitly released. For example, releasing a certain experiment has the effect of releasing all its runs as well.

6.3 Versioning

SRA submissions are not explicitly versioned. Rather, a complete change history is stored for metadata and any version of the metadata can be accessed. Content such as run and analysis data are never modified. If these must be changed then current ones are deprecated (Withdrawn) and replacements added.

6.4 Curation

From time to time NCBI needs to update metadata in order to correct mistakes, propagate changes in other resources (for example taxonomic changes), and edit information in order to comply with editing requirements, copyrights, and data release policies. These "curation" changes may occur without necessarily seeking the approval of the original submitter. Run and analysis data will never be changed in this way. Also, original titles, descriptions, and names will be preserved as much as possible.

6.5 Withdraw

This command withdraws the specified object. If the object is an SRA submission, then all the components of that submission are withdrawn.

Withdrawal simply marks the record as deprecated. Withdrawn records are never deleted (except for technical reasons including for example a loading error). Withdrawn records can still be accessed by accession, but the accession will be marked as having been withdrawn.

7 Examples

7.1 Microbial Whole Genome Sequencing Project

The study *Methylobacterium extorquens PA1 Whole Genome Sequencing Project* uses a single-ended random shotgun library of a microbial isolate sequenced on the 454 GS 20 platform, resulting in one experiment and four runs. See:

ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead/SRA000110

7.2 Epigenetics Study

The study *High-Resolution Profiling of Histone Methylations in the Human Genome* is a ChIP-Seq study performed on the Illumina 1G platform, resulting in 23 experiments (one for each sample) and 76 runs. See:

ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead/SRA000206

7.3 Transcriptome Study

The study *Manduca sexta ESTs under varying conditions* sequences an EST library on the 454 GS 20 platform with a short flow count configuration, resulting in one experiment and one run. See:

ftp://ftp.ncbi.nih.gov/pub/TraceDB/ShortRead/SRA000208